DRAFT SF 298

1. Report Date (dd-mm-yy) 29 June 1999	2. Report Type Final	3. Dates covered (from... to) Jan 99 - Jun 99
4. Title & subtitle Web Enabled DROLS Verity TopicSets Final Report		5a. Contract or Grant # SP4700-99-M-0211
		5b. Program Element #
6. Author(s) Tong, Richard M. Appelbaum, Lee		5c. Project #
		5d. Task #
		5e. Work Unit #
7. Performing Organization Name & Address Tarragon Consulting Corporation 1563 Solano Avenue, #350 Berkley, California 94707		8. Performing Organization Report #
9. Sponsoring/Monitoring Agency Name & Address Defense Technical Information Center 8725 John J. Kingman Road Ft. Belvoir, Virginia 22060		10. Monitor Acronym
		11. Monitor Report #

12. Distribution/Availability Statement
Approved for Public Release

13. Supplementary Notes

14. Abstract
The focus of this effort has been the design and development of automatically generated TopicSets and HTML pages that provide the basis of the required search and browsing capability for DTIC's Web Enabled DROLS System. The Tarragon developed components of the system are described in this Final Report. A key freature of these components is that they are maintainable and extensible by DTIC personnel using a suite of software utilities also provided by Tarragon as part of this effort.

15. Subject Terms
Search Engines, DROLS, Hierarchy Search Feature

Security Classification of			19. Limitation of Abstract	20. # of Pages	21. Responsible Person (Name and Telephone #)
16. Report U	17. Abstract U	18. This Page U	Unlimited		Jeff Davidson (703/767-7043)

WEB ENABLED DROLS

VERITY TOPICSETS

◆

FINAL REPORT

JUNE 30, 1999

Order No. SP4700-99-M-0211
DEFENSE TECHNICAL INFORMATION CENTER
8725 JOHN J. KINGMAN ROAD
FORT BELVOIR, VA 22060

◆ *Tarragon Consulting Corporation*
1563 SOLANO AVENUE, #350, BERKELEY, CA 94707
TEL: 510.526.7991 • FAX: 510.524.5622 • INFO@TCC.COM • WWW.TCC.COM

Table of Contents

1 Introduction

This document is the Final Report for the Verity TopicSets and associated search and browsing capabilities developed by Tarragon Consulting Corporation (Tarragon) for the Defense Technical Information Center (DTIC) under Order Number SP4700-99-M-0211.

The focus of this effort has been on the design and development of automatically generated TopicSets and HTML pages that provide the basis of the required search and browsing capability for DTIC's Web-Enabled DROLS System. The Tarragon developed components of the system are described in this Final Report. A key feature of these components is that they are maintainable and extensible by DTIC personnel using a suite of software utilities also provided by Tarragon as part of this effort. The Appendix to this report contains a complete set of instructions for creating and installing the Tarragon developed components.

1.1 Scope and Purpose

Tarragon has provided DTIC with a web-based "search hierarchy" and "hierarchy browsing" capability that exploits the current DTIC Thesaurus[1] and Corporate Source Hierarchy[2] and utilizes COTS knowledge management software from Verity, Inc. This capability is compatible with the three-tier architecture adopted by DTIC for its Web-Enabled DROLS system.

The successful completion of this effort has: (1) shown how the intellectual capital invested in the DTIC Thesaurus and Corporate Source Hierarchy can be exploited in a web-based environment that uses COTS software products; (2) demonstrated user-interface mechanisms for interacting with the thesaurus and source hierarchy in the web-based environment; and, (3) allowed DTIC to deploy a prototype Web-Enabled DROLS system that meets all the identified requirements.

1.2 Applicable Documents

The following DTIC document is applicable:

- Web Enabled DROLS/Verity TopicSets. Statement of Work. 7 January, 1999.

The following Tarragon documents are applicable:

- Web Enabled DROLS/Verity TopicSets. Requirements Specification (Version 1.1), March 31, 1999.
- Web Enabled DROLS/Verity TopicSets. Design Specification (Version 1.2), June 7, 1999.

1. *DTIC Thesaurus.* AD-A321 038, Oct 96.
2. *Source Hierarchy List. Volumes 1, 2, and 3.* AD-A281 100, AD-A281 101, and AD-A281 102. Jul 94.

2 Web Enabled DROLS Architecture

The web-based "search hierarchy" and "hierarchy browsing" capability provided by Tarragon is compatible with the DTIC architecture for a Web-Enabled DROLS shown in Figure 1. This

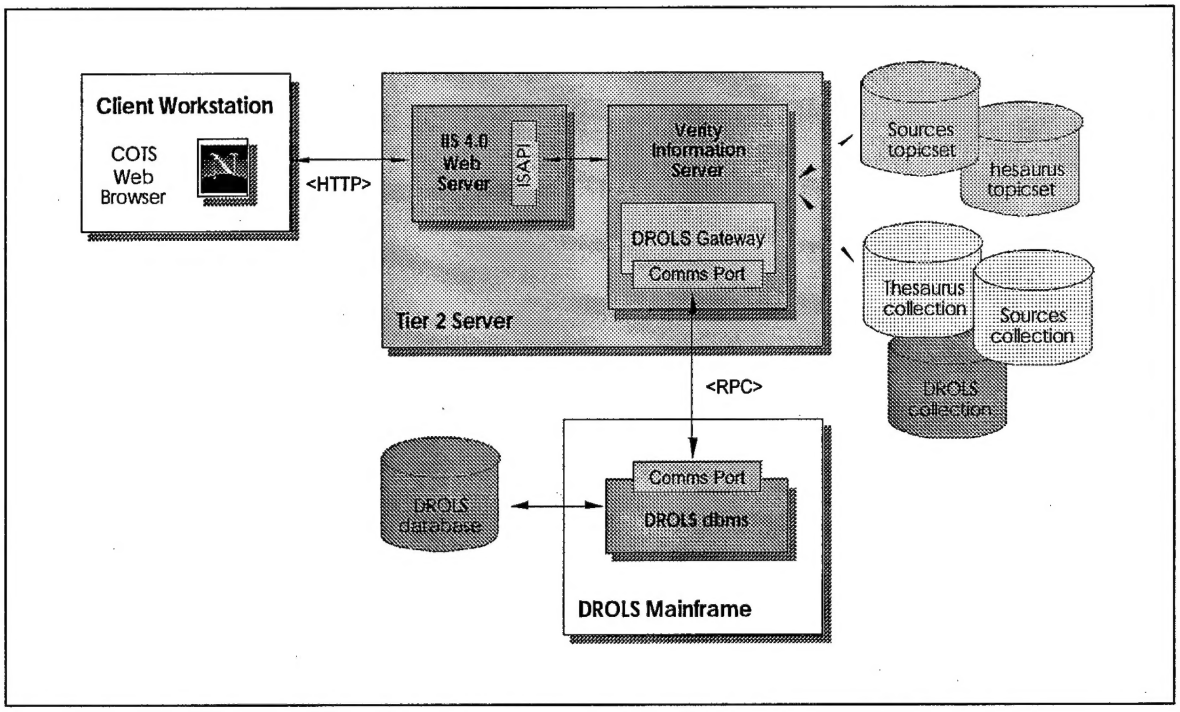

Figure 1: Three-Tier Web-Enabled DROLS Architecture

architecture uses Verity's Information Server product in a three-tier configuration in which Information Server communicates with the IIS Web Server using the ISAPI interface, and with the DROLS mainframe via a "DROLS Gateway" that itself uses RPCs as the base-level protocol.[3]

Information Server makes use of two sets of Tier-2 resources to provide the necessary search and retrieval capability—collections and topicsets. The DROLS Collection is one that Information Server builds from the DROLS records held in the DROLS dbms.[4] TopicSets are the data structures that Tarragon has developed to implement the DTIC Thesaurus and Corporate

3. DTIC has developed this custom gateway under a separate effort with Verity, Inc. This is documented in the Verity Professional Services Consulting Report, "DTIC November 23–December 4, 1999 (Revised 12/4/98)," by Jeremy Buckley.

4. The Verity collection is actually built against a custom citation format that is configurable via the DROLS Gateway. This gives DTIC the ability to easily modify and/or extended the range of information that is searchable. Note that this "indexing citation" is distinct from the "display citation" that is presented to the user once a DROLS record is to be viewed. (This causes a slight complication in the TopicSet design for the thesaurus—see Section 3.)

Source Hierarchy in a way that they can be used directly by Information Server. The key ideas behind the design of TopicSets are reviewed in Section 3 and Section 4.

In addition, Tarragon has developed two special Verity collections of HTML pages that support search and browsing of the contents of the Thesaurus and the Corporate Source Hierarchy. The key ideas behind the design of these collections are reviewed in Section 5 and Section 6.

3 TopicSet for the DTIC Thesaurus

The TopicSet designed and developed for the DTIC Thesaurus enables end-users to perform both simple and hierarchical descriptor searches over the DROLS database. The TopicSet is automatically generated from the DTIC Thesaurus and makes some assumptions about the format of both the files that contain the thesaurus data and the format of the citations used to build the Verity collection.

The software utilities provided by Tarragon assume that the DTIC Thesaurus is available in a data file that has the following format:

```
SAMPLERS
 UF  IMPINGERS
 UF  SAMPLING APPARATUS
SAMPLING
 NT  CORE SAMPLING
 NT  OCEAN BOTTOM SAMPLING
 BT *COLLECTING METHODS
SAMPLING APPARATUS
 use SAMPLERS
SAN DIEGO BAY
 BT *BAYS
 BT *CALIFORNIA
SAN FRANCISCO BAY
 BT *BAYS
 BT *CALIFORNIA
SAND
 UFC  SANDBLASTING
 NT  GRIT
 BT  ORES(NONMETALLIC)
 BT  SOILS
SAND CASTING
 BT *CASTING
SAND FLEAS
 BT *PULEX
SANDBAGS
  BAGS FILLED WITH SAND; AS ONE USED IN A PILE TO FORM A
  WALL, A REVETMENT, A FIELD FORTIFICATION OR AS A
  PROTECTION FOR BUILDINGS.
```

In this format, each descriptor is on a line by itself and is followed by additional terms and possibly by a scope note. Tarragon's software utilities read this file and convert it into a sequence of individual topic definitions in Verity's "outline format." This outline format file can then be converted into a compiled TopicSet using a standard Verity software utility.[5]

The software utilities provided by Tarragon also assume that the DROLS collection has been built using citations in the following format:[6]

```
<HTML>
<HEAD>
<TITLE>
Elastic, Piezoelectric, and Dielectric Constants of Bi12GeO20
</TITLE>
</HEAD>
<BODY>
...
<TI>Elastic, Piezoelectric, and Dielectric Constants of
Bi12GeO20</TI>
...
<SC>011800</SC>
<CA>AIR FORCE CAMBRIDGE RESEARCH LABS L G HANSCOM FIELD MASS</CA>
...
<DE>
<WDE>*Bismuth compounds ZBismuth compoundsZ, </WDE>
Germanium compounds ZGermanium compoundsZ, Oxides ZOxidesZ,
Elastic properties ZElastic propertiesZ, Piezoelectricity
ZPiezoelectricityZ, Dielectric properties ZDielectric prop-
ertiesZ, Test methods ZTest methodsZ
</DE>
<ID>Bismuth germanium oxides</ID>
<AB>In order to resolve conflicting published results con-
cerning the values of the elastic, piezoelectric, and
dielectric constants of Bi12GeO20, an exhaustive study of
this material has been made. Over 45 samples of material
grown by three independent sources were tested. The piezo-
electric constant was determined using three different tech-
niques. In addition, the first accurate measurements of the
dielectric constant of Bi12GeO20 at microwave frequencies
have been made.  (Modified author abstract)
</AB>
</BODY></HTML>
```

5. Building the complete set of topics that correspond to the DTIC Thesaurus takes approximately one hour. The bulk of this time is used in processing the data file that contains the descriptors; less than 10% of the time is used in compiling the TopicSet from the outline file.

6. Some of the details have been elided to emphasize the key elements in the citation as they relate to the construction of the TopicSet. Some HTML formatting tags have also been removed for clarity.

Note that the citation is an HTML page and that it contains a number of key "fields" that are used in the construction of the collection (e.g., the title field indicated by the <TI></TI> tag pair, the source code field <SC></SC>, the corporate author field <CA></CA>, etc.). In particular there is a descriptors field, indicated by the <DE></DE> tag pair, which has embedded in it a weighted descriptors field, indicated by the <WDE></WDE> tag pair.

The descriptors field contains two forms of the descriptor—a form that is identical to the one found in the thesaurus (e.g., Bismuth compounds) and a form that surrounds the thesaurus form with a pair of Z's (e.g., ZBismuth compoundsZ). The reason for including this Z-form is that when search is performed we want to avoid matching on individual words in a descriptor when it actually a sequence of terms. So when searching for "Bismuth" we only want to match on Bismuth and not on Bismuth compounds. By enclosing the descriptor in a pair of Z's we can ensure that a search specifically for the term "Bismuth" becomes a search for "ZBismuthZ" and does not therefore match the fragment "Zbismuth" in the phrasal form.

Given these assumptions, the software algorithms provided by Tarragon generate a topic that will search the collection for citations that contain the descriptor. So, to use another example, the descriptor MILITARY FORCES(UNITED STATES) defined by:

```
MILITARY FORCES(UNITED STATES)
 UF  ARMED FORCES(UNITED STATES)
 NT *AIR FORCE
 NT *ARMY
 NT  COAST GUARD
 NT  MARINE CORPS
 NT  NAVY
 NT *SPECIAL OPERATIONS FORCES
 BT *MILITARY ORGANIZATIONS
```

becomes the following topic:

```
_MILITARY_FORCES_UNITED_STATES <Or>
* _MILITARY_FORCES_UNITED_STATES_IN_ZONES <Or>
** 1.00 _MILITARY_FORCES_UNITED_STATES_IN_WDE <In>
        /Zonespec = "WDE"
*** "ZMILITARY FORCES ( UNITED STATES ) Z"
** 0.00 _MILITARY_FORCES_UNITED_STATES_IN_WDE1 <In>
        /Zonespec = "WDE"
*** "MILITARY FORCES ( UNITED STATES ) "
** 0.50 _MILITARY_FORCES_UNITED_STATES_IN_DE <In>
        /Zonespec = "DE"
*** "ZMILITARY FORCES ( UNITED STATES ) Z"
** 0.00 _MILITARY_FORCES_UNITED_STATES_IN_DE1 <In>
        /Zonespec = "DE"
*** "MILITARY FORCES ( UNITED STATES ) "
* _AIR_FORCE
* _ARMY
* _COAST_GUARD
```

```
* _MARINE_CORPS
* _NAVY
* _SPECIAL_OPERATIONS_FORCES
```

There are several things to note about this topic. It uses the Verity "outline format" in which asterisks (*) are indicators of hierarchy. So the top-level topic is `_Military_Forces_United_States`[7] and its immediate children are those items preceded by a single asterisk. Notice too that the topic structure has two types of children—specifications for actual search (the two- and three-asterisk items), and other topics which correspond to the narrower terms of the top-level descriptor topic. The structure is illustrated in Figure 2.

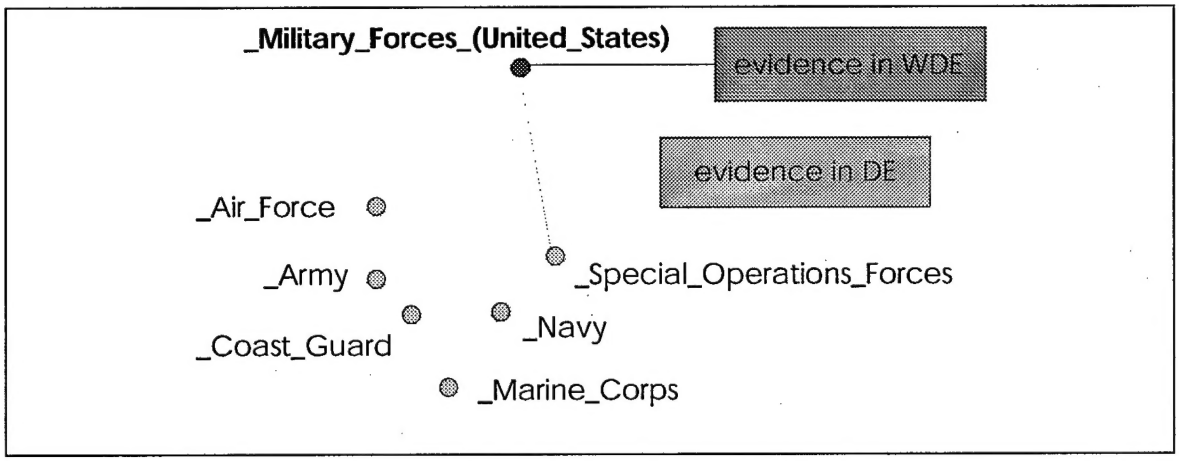

Figure 2: Structure of a Descriptor Topic

Since each of the topics that correspond to other descriptors are defined using a similar structure, the whole network of topics (the TopicSet) is both a representation of the DTIC Thesaurus and a specification for how to search for descriptors in the DROLS collection.

The weighted branches in the topic definition specify the relative importance of the search patterns (the "evidence" in Verity terminology) in determining whether the citation matches the search specification. In these topics, weights are just used to distinguish between matching descriptors in the WDE field (or "zone" in Verity terminology) and the DE field. Matches in the WDE field get a weight of 1.00; matches in the DE field get a weight of 0.50.[8]

Another feature of the topic definition is the presence of zero weighted branches. These are included to ensure proper highlighting behavior with the display form of the citation. They have no impact on the search behavior.

---

7. This leading underbar notation is a Tarragon convention for naming topics. It helps ensure uniqueness of topic names and make them clearly distinct from text phrases that might also be part of a topic definition.

8. In the current version of the Web Enabled DROLS, the weights are used to sort the retrieval results. Citations with a match in the WDE field appear before citations with a match in the DE field.

Finally, the topic is structured so that it will support both simple and hierarchical search. If a simple descriptor search is requested, then only the topic branch _MILITARY_FORCES_UNITED_STATES_IN_ZONES is used; if a hierarchical descriptor search is requested, then the complete topic _MILITARY_FORCES_UNITED_STATES is used.[9]

4 TopicSet for the DTIC Corporate Source Hierarchy

The TopicSet designed and developed for the DTIC Corporate Source Hierarchy enables end users to perform both simple and hierarchical source code searches over the DROLS database. As with the Thesaurus TopicSet, the TopicSet for sources is automatically generated from the DTIC Corporate Source Hierarchy.[10]

The software utilities provided by Tarragon assume the same indexing citation format as before, as well as a sources data file that has the following format:

```
NAVAL OCEAN SYSTEMS CENTER  SAN DIEGO CA          393159
 06 NAVAL ELECTRONICS LAB CENTER                  403940
    SAN DIEGO CA
  07 NAVAL ELECTRONICS LAB CENTER FOR             403032
     COMMAND CONTROL AND COMMUNICATIONS
     SAN DIEGO CA
  07 NAVY ELECTRONICS LAB  SAN DIEGO CA           253550
   08 NAVY ELECTRONICS LAB  SAN DIEGO CA          400133
      RADIO PHYSICS DIV
 06 NAVAL OCEAN SYSTEMS CENTER  DAHLGREN VA       396082
 06 NAVAL OCEAN SYSTEMS CENTER  KAILUA HI         393845
 06 NAVAL OCEAN SYSTEMS CENTER  SAN DIEGO CA      418181
    BLOCK PROGRAM
 06 NAVAL OCEAN SYSTEMS CENTER  SAN DIEGO CA      393741
    VARIOUS/MULTIPLE SPONSORED WORK
 06 NAVAL SHORE ELECTRONICS ENGINEERING           393247
    ACTIVITY PACIFIC
    FPO SAN FRANCISCO 96610
 06 NAVAL UNDERSEA CENTER  SAN DIEGO CA           390458
  07 NAVAL UNDERSEA CENTER  KAILUA HI             390379
     HAWAII LAB
  07 NAVAL UNDERSEA CENTER  PASADENA CA           390591
  07 NAVAL UNDERSEA RESEARCH AND DEVELOPMENT      405998
     CENTER  PASADENA CA
   08 NAVAL UNDERSEA RESEARCH AND DEVELOPMENT     389481
```

9. In the deployed Web Enabled DROLS system there is a layer of software between the User Interface and the invocation of the TopicSet that allow the system to select the correct topic based on the user's inputs.

10. Building the complete set of topics that correspond to the DTIC Corporate Source Hierarchy takes approximately one hour. The bulk of this time is used in processing the data file that contains the corporate sources; less than 10% of the time is used in compiling the TopicSet from the outline file.

```
   CENTER  POINT MUGU CA
   MARINE BIOSCIENCE FACILITY
08 NAVAL UNDERSEA RESEARCH AND DEVELOPMENT      404762
   CENTER  SAN DIEGO CA
 09 NAVAL UNDERSEA RESEARCH AND DEVELOPMENT     406558
    CENTER  SAN DIEGO CA
    ARCTIC SUBMARINE LAB
 09 NAVAL UNDERSEA WARFARE CENTER               403023
    SAN DIEGO CA
  10 NAVAL UNDERSEA WARFARE ENGINEERING         433843
     STATION  SAN DIEGO CA  SOUTHERN
     CALIFORNIA DETACHMENT
08 NAVAL UNDERSEA WARFARE CENTER                403369
   PASADENA CA
 09 NAVAL UNDERSEA WARFARE CENTER               389086
    KAILUA HI
```

Since source codes are unique identifiers of corporate authors, the resulting topics have a much simpler structure than those for the descriptors. In particular, there is no need for special forms of the source code or additional topic branches to ensure proper highlighting. Thus the corresponding topic for "NAVAL OCEAN SYSTEMS CENTER SAN DIEGO CA" is:

```
_393159 <Or>
* _393159_In_Zone <In>
  /Zonespec = "SC"
** "393159"
* _403940
* _396082
* _393845
* _418181
* _393741
* _393247
* _390458
```

which use the same Verity "outline format" and in which the topic names are based on the source code for the corporate author.

Note that no weights are required here—the source code is either matched or it is not; and that there is only one "evidence" branch in the topic—all the others are topics for children (_403940, _396082, ...) that are similarly defined. The structure of this topic is illustrated in Figure 3. As with the thesaurus, the whole network of topics is both a representation of the DTIC Corporate Source Hierarchy and a specification for how to search for source codes in the DROLS collection.

Each source topic is structured so that it will support both simple and hierarchical search. If a simple source code search is requested, then only the topic _393159_In_Zone is used; but if a hierarchical source code search is requested, then the complete topic _39315 is used.

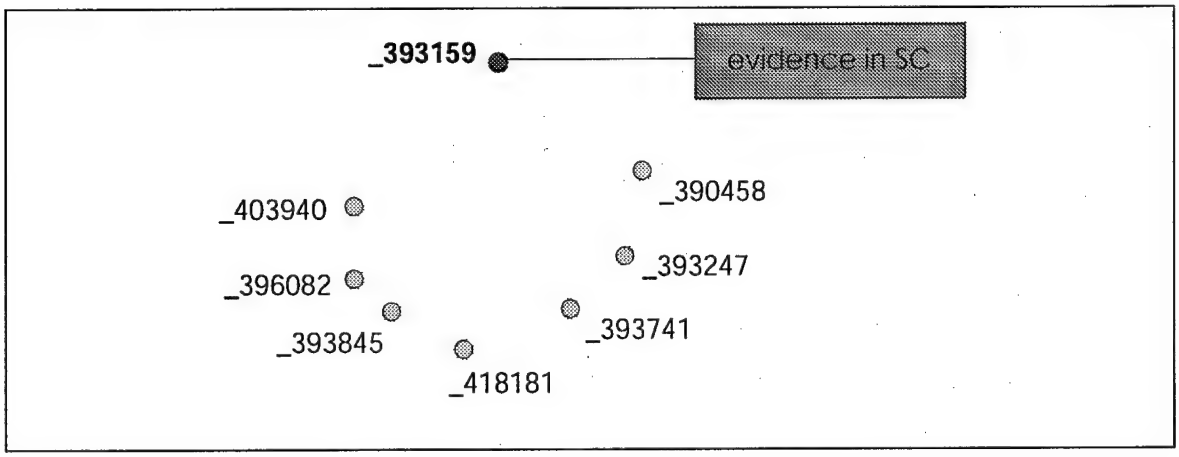

Figure 3: Structure of a Corporate Author Topic

5 Browsing the DTIC Thesaurus

To provide an end-user capability to browse and search the DTIC Thesaurus, Tarragon designed and developed a strategy that involves the following steps:[11]

1. For each descriptor, automatically generate an HTML page that contains information about the descriptor and its associated terms (i.e., narrower terms, broader terms, ...), and make each term on the page a hyperlink.
2. Use Information Server to index these pages and create a searchable Verity collection
3. Provide a search and browsing interface that can exploit the Verity collection and the Verity query language.

This strategy has a number of advantages;

1. It makes use of straightforward HTML avoiding the need for users to have web browsers that support advanced features such as Java or style sheets.
2. It leverages DTIC's investment in Verity software. The same engine used to index and search the DROLS citation is used to index and search the DTIC Thesaurus.
3. The software utilities provided by Tarragon enable to DTIC to easily modify and customize the "look and feel" of the pages and the ways in which they are indexed.

Figure 4 illustrates how a descriptor entry in the DTIC Thesaurus is converted into an HTML page.[12] They key idea is that each reference to another descriptor on the page becomes a hyperlink (i.e., is embedded in `<A></A>` HTML tags) to the HTML page for that descriptor. The figure

11. The first two steps use software utilities developed by Tarragon that run as a batch process and take approximate 30 minutes to generate the set of HTML pages and then index them. The pages themselves require approximately 500Mb of storage and the resulting Verity index requires approximately 50Mb.

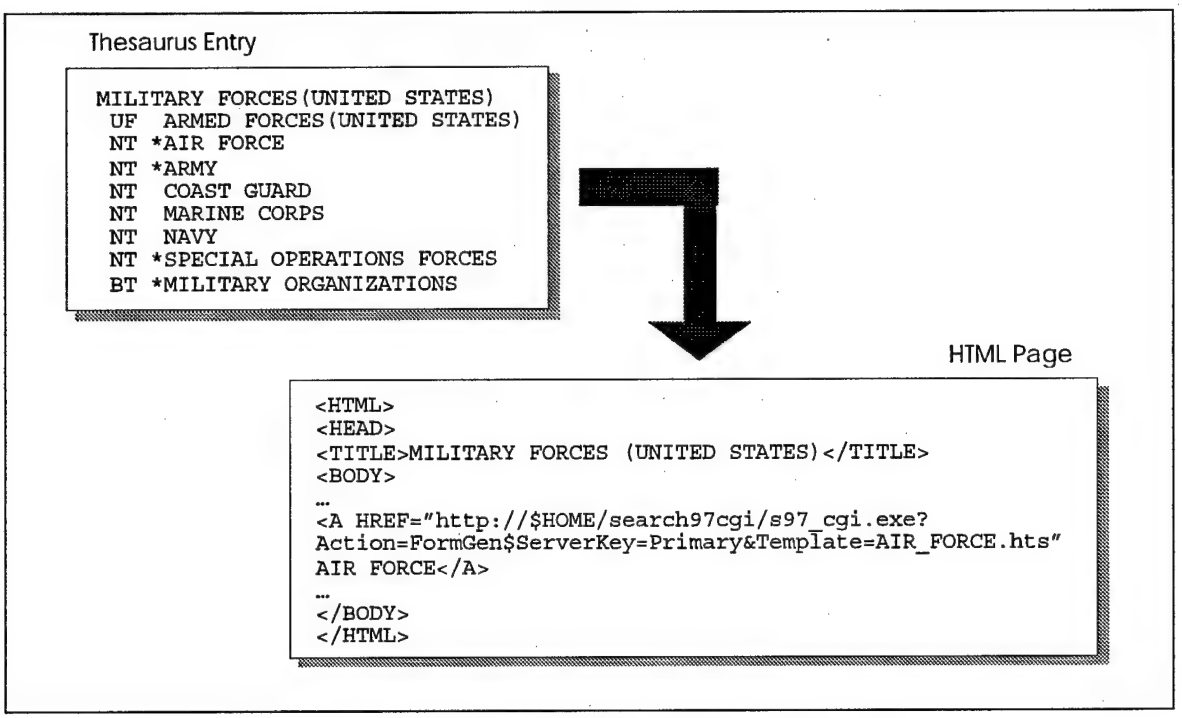

Figure 4: Illustration of Thesaurus Page Construction

shows just a few elements of the HTML page for the descriptor `MILITARY FORCES (UNITED STATES)`.

Note that the narrower term `AIR FORCE` becomes a link in the page that points to the HTML page that contains the information about this term (i.e., `AIR_FORCE.hts`). All the other narrower terms (`ARMY`, `COAST GUARD`, `MARINE CORPS`, `NAVY`, `SPECIAL OPERATIONS FORCES`) are treated in the same way.

The complete set of HTML pages that result from this process are interconnected and can be browsed by following the various links. At the same time, since they are indexed by Information Server, they are also directly searchable thus allowing users to quickly and easily find the thesaurus entries they need. To facilitate accurate search each HTML page contains a descriptor field (a "zone" in Verity terminology) that contains a searchable form of the descriptor. In this example it is:

```
<DE>MILITARY FORCES ( UNITED STATES ) </DE>
```

12. The Design Specification describes these pages in more detail. We show here just a fragment in order to illustrate the process.

so that all of the capabilities in the Verity search language can be applied. Thus end users can invoke the various wildcarding and proximity search operators that Verity supplies to locate a descriptor when only a part of the descriptor name is known.[13]

6 Browsing the DTIC Corporate Source Hierarchy

To provide an end-user capability to browse and search the DTIC Corporate Source Hierarchy, Tarragon used exactly the same strategy as that developed for the DTIC Thesaurus. Thus the equivalent process is:[14]

1. For each corporate author, automatically generate an HTML page that contains information about the corporate author and its associated entries (i.e., child organizations, parent organization, Source Code, ...), and make each entry on the page a hyperlink.
2. Use Information Server to index these pages and create a searchable Verity collection
3. Provide a search and browsing interface that can exploit the Verity collection.

The complete set of HTML pages that result from this process are interconnected and can be browsed by following the various links. At the same time, since they are indexed by Information Server, they are also directly searchable thus allowing users to quickly and easily find the corporate author they need. As with the thesaurus pages, each corporate author page contains a corporate author field that supports a full Verity search. In the example, this would be:

```
<CAU>NAVAL OCEAN SYSTEMS CENTER SAN DIEGO CA</CAU>
```

allowing end-users to locate corporate authors when only a portion of the name is known.

7 Test and Evaluation

Tarragon performed a series of test and evaluation exercises to assess the accuracy and effectiveness of the "search hierarchy" and "hierarchy browsing" functionality described above.

To test the search hierarchy, Tarragon used a sample set of 1000 DROLS citations provided by DTIC along with 10 test queries and their "ground truths" (i.e., the set of hits produced by these queries against the standard DROLS database restricted to these same 1000 citations). In all cases, the TopicSet version of the query produced exactly the same result as the standard DROLS query.

13. As with the other components that Tarragon has developed on this effort, the user interfaces are being developed by DTIC. What the end-user actually sees and which search operators are made available is a DTIC decision.

14. The first two steps use software utilities developed by Tarragon that run as a batch process and take approximate an hour to generate the set of HTML pages and then index them. The pages themselves require approximately 1Gb of storage and the resulting Verity index requires approximately 50Mb.

To test the hierarchy browsing capability, Tarragon developed prototype user interfaces that were used by DTIC personnel to assess the utility of the browsing capability. Both individual and collective feedback was used to refine the design of the HTML pages that constitute the basis for the Thesaurus and Corporate Source Hierarchy browsing and searching capability.

By involving several groups of DTIC personnel, and by providing early "hands-on" access to the search and browsing capabilities, the design for the search and browsing capabilities was extensively tested and reviewed in a variety of contexts. The final design and implementation thus reflects a consensus on the best strategy for providing these capabilities in the operational version of the Web-Enables DROLS system.

8 Summary and Conclusions

The successful completion of the effort described in this report has shown how the intellectual capital invested in the DTIC Thesaurus and DTIC Corporate Source Hierarchy can be exploited in a web-based environment that uses standard commercial document retrieval software, and has demonstrated effective user-interface mechanisms for interacting with the thesaurus and source hierarchy in a standard, widely-available web-based environment.

The project was completed on time and met all the requirements identified. The key factors in the success of this project are:

- The overall requirements and design for the search and browsing capability were developed jointly by Tarragon and DTIC. The use of early prototypes and the input from several DTIC constituencies, ensured that the final implementation of these capabilities meets real needs and provides high-value to DTIC's end-users.
- The same Verity software used to provide search of the DROLS citations is used to provide search of the DTIC Thesaurus and DTIC Corporate Source Hierarchy. This leverages the investment made by DTIC, and simplifies the long-term maintenance issues.
- All user interaction, both searching and browsing, is via standard HTML pages and commercial web-browsers. This ensures the widest range of access, and eliminated the need for complex client-side software application development.
- The software utilities developed by Tarragon that construct the TopicSets and the Collections for the Thesaurus and Corporate Source Hierarchy are provided in source code form, thus allowing DTIC to modify these as they need to meet any changing system requirements.
- The use of a configurable gateway to the DROLS mainframe allows DTIC to easily control the format of citations and thus provide a variety of formats that can be tailored as necessary to meet future search and browsing needs.

Given the flexible design and the demonstrated capability to access DTIC data using a high-performance retrieval engine, the Web-Enabled DROLS platform and its components might be used in a number of ways to provide additional capabilities both within DTIC and to DTIC's end-user population. The following section outlines a number of possible directions that Tarragon believes would allow DTIC to capitalize on the investment it has made in current effort.

8.1 Future Developments

The current effort has produced several key knowledge resources that can be exploited and extended in additional ways to benefit DTIC and its users. The range of options is large, but some natural future developments would include:

Enhancing the User Search Experience.

Two possible ways of enhancing the search experience for the end-user are: (a) to provide category-based access to the DROLS data; and, (b) to provide topic-based search of additional information in the DROLS database.

Verity's Knowledge Organizer product would provide a direct way of implementing a category structure based on the Thesaurus TopicSet. This would provide a natural "Yahoo!-like" drill-down for the DROLS user without having to do any extra TopicSet development or document indexing. DTIC would also be able to create custom categories (i.e., "views") that reflect current subjects of interest, and the Knowledge Organizer framework also supports the idea of a "my-DTIC" capability that would allow end-users to create personalized structures for accessing the DROLS data.

The current TopicSets search over only a limited set of information in the DTIC database. Extending them to search over additional fields, such as the Abstract, Identifiers and Title, would provide users with the ability to perform more comprehensive and flexible searches.

The TopicSets could also be extended to search over the full-text of those Technical Reports that are available in this form. By adding a stronger semantic model than the Thesaurus currently supports, the TopicSets could become a basis for standard information finding "ontology" that could be used for a variety of technical analysis problems.

DTIC as a DoD Portal.

The concept of an "Enterprise Information Portal" is rapidly gaining support in the commercial arena as a way of providing users with seamless, uniform access to all data of interest. Such a concept could be developed for DTIC using extensions of the knowledge resources developed on this effort.

The objective would be to make DTIC a DoD Portal giving users access not only to DTIC's own data but also to other Government resources. The TopicSet model could be extended to incorporate terms and expressions that would search over non-DROLS databases—such as the IACs, and those provided by NASA and DOE. The DTIC Thesaurus would become the standard search vocabulary so that users would need to become familiar with just one search metaphor in order to gain access to a wide range of resources.

The DTIC Portal would become a primary resource for DoD users who wish to find technical information. By combining this heterogeneous access capability with the enhanced search features described in the previous section, and with some of the advanced delivery options suggested in the following section, the DTIC Portal would appeal to a very wide range of users and needs.

Advanced Information Delivery Capabilities.

The TopicSets delivered on this effort could be combined with other Verity products to create additional information delivery capabilities for DTIC users. These include:

(a) A current awareness system using personal profiles based on the TopicSets and the Verity Agent Server. Users would define their interests in terms of the Thesaurus and Sources topics as well as other constraints (e.g., date, source, language), and the Verity Agent Server would deliver the information in whatever form required—as HTML pages, as email, as faxes, etc. This capability would become the DTIC equivalent of those many clipping and alerting services offered in the commercial marketplace.

(b) A CD-ROM based version of the "search hierarchy" and "hierarchy browsing" capability that utilizes Verity's CD Publisher product and that would give users local access to the DROLS data with the option of web-based connectivity back to DROLS to get updates and the most recent information. The CD Publisher product would allow DTIC to ship a complete search and browsing capability to users, but also allow them to get the most recent information over the web thus avoiding the "information staleness" problem associated with cutting CDs.

(c) Combining the TopicSets with Verity Spider to provide category-based access to DoD websites of interest to DTIC users. The Verity Spider is a fully-functional, high-performance indexing tool that can be configured to crawl both the open and internal networks. By combining it with extended versions of the Thesaurus and Sources TopicSets, a highly-focused index of web pages could be developed that could be integrated with the category-based access capability or with the DTIC Portal concept described in previous sections.

A TopicSet and Collection Generation Guide

This appendix contains a complete set of instructions for generating the Thesaurus and Corporate Source Hierarchy TopicSets, and the Thesaurus and Corporate Source Hierarchy collections needed by the Verity Information Server as described in the main body of this report.

The first section contains a description of the installation directories, the remaining sections contain detailed instructions for creating the TopicSets and collections.

A.1 Directory Mapping

The following delineates the location on the DTIC ClearPath NT platform (CPNT) of all the files used to build the HTML browser pages, supporting Verity collections, and the topic outline files for the descriptor hierarchy and corporate source hierarchy features.

F:\Perl

Contains the Perl Compiler

1. The compiler `perl.exe`
2. The Perl Dynamic Load Library (DLL) `perl.dll`

F:\Utilities

Contains Perl scripts to build topic outline files and search forms (HTML pages) for sources and descriptors

1. `mk_descriptor_pages.pl` creates an HTML page for each descriptor, with links to related pages, providing a browser for the DTIC thesaurus
2. `mk_descriptor_topics.pl` creates topic outline files for the descriptors
3. `mk_sources_pages.pl` creates an HTML page for each corporate author, with links to related pages, providing a browser for the DTIC sources
4. `mk_sc_topics.pl` creates topic outline files for the source codes

F:\Descriptors

Contains the Descriptor "documents":

1. Copy of the `postings.txt` input file
2. HTML pages for descriptors from the DTIC controlled vocabulary
   - One page for each descriptor
   - About 16,500 files

F:\Sources

Contains the Corporate Source "documents":

1. Copy of the `sources.txt` input file
2. HTML page for corporate authors identified by unique source codes
   - One page for each source code
   - About 38,850 files

F:\Styles

Contains the Verity style files for the hierarchy related collections:

1. Style files for the Descriptors Collection are in the `descriptors` sub-directory
2. Style files for the Sources Collection are in the `sources` sub-directory

F:\colls

Contains Verity Collections for descriptors and sources:

1. Descriptors collection sub-directory: `Descriptors`
2. Sources collection sub-directory: `Sources`

F:\Topics

Contains specifications for the Verity TopicSets:

1. `DescriptorOutlines` contains outline (.otl) files for descriptors
2. `SC_Outlines` contains outline (.otl) files for sources
3. `mktopicsX.bat`—batch file used to compile the descriptor outline files
4. Compiled topic sets (result of running `mktopicsX.bat`)

A.2 Creating the Descriptor Topics

Follow these steps to create the Verity Topic outline files (.otl) for the descriptors. We need multiple files because there are too many topics for the Verity `mktopics` utility to process in a single file. As a result, many topics are defined in terms of topics that are themselves defined in another outline file. Topics with these "external references" may and probably will cause `mktopics` to print warnings. However, these warnings may be ignored because the Verity engine links the topics during retrieval.

Input Files

1. `postings.txt` contains the input data from the DTIC Thesaurus as maintained in the Lexico system. It is kept in the `F:\Utilities` directory.
2. `mk_descriptor_topics.pl` is the Perl program needed. It is also kept in the `F:\Utilities` directory.
3. `mktopicsX.bat` is a DOS based batch command file containing the DOS commands to (re-)generate all the necessary Topicsets.

Output Files

1. Topic outline files, defined in the same directory from which the program is executed. These are intermediate files that are input into the Verity topic definition utility `mktopics`.
2. Verity Topic files in the directories `desc1`, `desc2`, `desc3`, and `desc4`.

Verity Outline File Creation Process

From Windows NT Command Prompt:

1. Make `F:\topics\DescriptorOutines` the current directory.
2. Empty the directory.
3. Copy the Perl program `mk_descriptor-topics.pl` from the `F:\Utilities` directory.
4. Copy the `postings.txt` input file from the `F:\Utilities` directory.
5. Execute the command:

```
F:\perl\perl.exe mk_descriptor_topics.pl postings.txt
```

Verity TopicSet Creation Process

1. Make `F:\topics` the current directory.
2. Run the batch file `mktopicsX.bat`
3. Verify that the file `C:\Verity\IS\common\kbases.kbm` includes references to the four Topicset Directories created in the `Topics` directory .

Example:

```
F:
cd \
cd \topics\DescriptorOutlines
copy F:\Utilities\mk_desciptor_topics.pl
```

```
copy F:\Utilities\postings.txt
F:\Perl\perl.exe mk_descriptor_topics.pl postings.txt
cd ..
F:\Utilities\mktopics_descriptors.bat
```

A.3 Creating the Corporate Source Topics

Follow these steps to create the outline files for the source codes. We need multiple files because there are too many topics for a single file. As a result, many topics are defined in terms of topics that are defined in another outline file. Topics with these "external references" cause mktopics to print warnings; however, the Verity engine links the topics during retrieval.

Nota Bene: The Corporate Source Browser Pages must be created before the Corporate Source Topics are created to provide the correct input file to this procedure.

Input Files

1. `sources.txt` contains the input data from the DTIC Corporate Source Authority System. It is kept in the `F:\Utilities` directory.
2. `mk_sc_topics.pl` is the Perl program needed. It is also kept in the `F:\Utilities` directory.

Output Files

1. Topic outline files, defined in the same directory from which the program is executed. These are intermediate files that are input into the Verity topic definition utility.
2. Verity Topic files in the directories `sc1`, `sc2`, `sc3`, and `sc4`.

Verity Outline File Creation Process

From Windows NT Command Prompt:

1. Make `F:\topics\SourceOutines` the current directory.
2. Empty the directory.
3. Copy the Perl program `mk_sc_topics.pl` from the `F:\Utilities` directory.
4. Copy the `sources.txt` input file from the `F:\Utilities` directory.
5. Execute the command:

```
F:\perl\perl.exe mk_sc_topics.pl sources.txt
```

Verity TopicSet Creation Process

From Windows NT Command Prompt:

1. Make `F:\topics` the current directory.
2. Run the Verity Utility `mktopics` for each outline file
3. Verify that the file `C:\Verity\IS\common\kbases.kbm` includes references to the four Topicset Directories created in the `F:\Topics` directory.

Example:

```
F:
cd \
cd \topics\SourceOutlines
copy F:\Utilities\mk_sc_topics.pl
copy F:\Sources\sources.txt
F:\Perl\perl.exe  mk_sc_topics.pl  sources.txt
cd ..
F:\Utilities\mktopics_sources.bat
```

Nota Bene: The topic sets (`sc1`, `sc2`, etc.) must be included in the `KBASES.kbm` file.

A.4 Creating the Descriptor Pages and Descriptor Collection

Follow these steps to create the HTML files for descriptor browsing and to create the `load.blk` file used to build the collection for these files. The pages are then indexed in a new Verity collection.

Input Files

1. `postings.txt` contains the input data from the DTIC Thesaurus as maintained in the Lexico system. The latest version is kept in the `F:\Utilities` directory.
2. `mk_descriptor_pages.pl` is the Perl program needed. It is also kept in the `F:\Utilities` directory.

Output Files

1. HTML files for the descriptors are generated in the same directory from which the program is executed.
2. `load.blk` used to build the collection from the HTML pages files is generated in the same directory also.

Descriptor Page Generation Process

From Windows NT Explorer:

1. Copy the new `postings.txt` file into the `F:\Utilities` directory.
2. Empty the directory `F:\Descriptors`
3. Make `F:\Descriptors` the current directory.
4. Copy the Perl program `mk_descriptor_pages.pl` from the `F:\Perl` directory.
5. Copy the `postings.txt` file from the `F:\Utilities` directory.
6. Open a Command Prompt NT window.
7. Make `F:\Descriptors` the current directory.
8. Execute the command:

```
F:\Perl\perl.exe mk_descriptor_pages.pl postings.txt
```

Descriptor Collection Re-Creation Process

From Windows NT Explorer:

1. Empty the directory `F:\colls\descriptors`
2. Open a Command Prompt NT window.
3. Make `F:\colls\descriptors` the current directory.
4. Create the new collection:

```
mkvdk -create -collection F:\colls\descriptors
      -style F:\Styles\Descriptors
```

Descriptor Collection Indexing Process

From a Command Prompt NT Window:

1. Make `F:\colls\descriptors` the current directory.
2. Execute a bulk load:

```
mkvdk -collection F:\colls\descriptors -bulk load.blk
```

Example:

```
F:
cd \
cd \descriptors
erase *.* (y)
copy F:\Utilities\mk_descriptor_pages.pl
```

```
copy F:\Utilities\postings.txt
F:\perl\perl.exe mk_descriptor_pages.pl postings.txt
cd ..\colls
rmdir /s descriptors (y)
cd ..
mkvdk -create -collection descriptors -style F:\Styles\descriptors
mkvdk -collection descriptors -bulk F:\descriptors\load.blk
```

Nota Bene: If the collection is new, then it must be imported into the Verity Information Server environment by using the Collection Manager in the Administrator's Tool.

A.5 Creating the Corporate Source Pages and Sources Collection

Follow these steps to create the HTML files for source browsing and to create the load.blk file used to build the collection for these files.

Input Files

1. `sources.txt` contains the input data from the DTIC Corporate Source Authority System as maintained on the Unisys Classified Enterprise Server. The latest version is kept in the `F:\Utilities` directory.
2. `mk_source_pages.pl` is the Perl program needed. It is also kept in the `F:\Utilities` directory.

Output Files

1. HTML files for the sources are generated in the same directory from which the program is executed.
2. `load.blk` used to build the collection from the HTML files is generated in the same directory also.

Source Page Generation Process

From Windows NT Explorer:

1. Copy the new `sources.txt` file into the `F:\Utilities` directory.
2. Empty the directory `F:\Sources`
3. Make `F:\Sources` the current directory.
4. Copy the Perl program `mk_source_pages.pl` from the `F:\Perl` directory.
5. Copy the `sources.txt` file from the `F:\Utilities` directory.
6. Open a Command Prompt NT window.
7. Make `F:\Sources` the current directory.

8. Execute the command:

```
F:\Perl\perl.exe mk_source_pages.pl sources.txt
```

Sources Collection Re-Creation Process

From Windows NT Explorer:

1. Empty the directory `F:\colls\sources`
2. Open a Command Prompt NT window.
3. Make `F:\colls\sources` the current directory.
4. Create the new collection:

```
mkvdk -create -collection F:\colls\sources
      -style F:\Styles\sources
```

Descriptor Collection Indexing Process

From a Command Prompt NT Window:

1. Make `F:\colls\sources` the current directory.
2. Execute a bulk load:

```
mkvdk -collection F:\colls\sources -bulk load.blk
```

Example:

```
F:
cd \
cd \sources
erase *.* (y)
copy F:\Utilities\mk_sources_pages.pl
copy F:\Utilities\sources.txt
F:\perl\perl.exe mk_source_pages.pl sources.txt
cd ..\colls
rmdir /s sources (y)
cd ..
mkvdk -create -collection sources -style F:\Styles\sources
mkvdk -collection sources -bulk F:\sources\load.blk
```

Nota Bene: If the collection is new, then it must be imported into the Verity Information Server environment by using the Collection Manager in the Administrator's Tool.